Data and Society
The Data-driven World – Lecture 2

1/28/21
Today (1/28/21)

• Personal Essay Assignment and Instructions
• Lecture 2 /Discussion
• Model Presentation
Reading for 2/1/21

• Read this article for the 2/1 class discussion.

• “Why It’s So Freaking Hard to Make a Good COVID-19 Model”, 538

<table>
<thead>
<tr>
<th>Date</th>
<th>Topic</th>
<th>Speaker</th>
<th>Date</th>
<th>Topic</th>
<th>Speaker</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-25</td>
<td>Introduction</td>
<td>Fran</td>
<td>1-28</td>
<td>The Data-driven World</td>
<td>Fran</td>
</tr>
<tr>
<td>2-1</td>
<td>Data and COVID-19</td>
<td>Fran</td>
<td>2-4</td>
<td>Data and Privacy -- Intro</td>
<td>Fran</td>
</tr>
<tr>
<td>2-8</td>
<td>Data and Privacy – Differential Privacy</td>
<td>Fran</td>
<td>2-11</td>
<td>Data and Privacy – Anonymity</td>
<td>Fran</td>
</tr>
<tr>
<td>2-15</td>
<td>NO CLASS / PRESIDENT’S DAY</td>
<td></td>
<td>2-18</td>
<td>Data and Privacy – Law</td>
<td>Ben Wizner</td>
</tr>
<tr>
<td>2-22</td>
<td>Digital rights in the EU and China</td>
<td>Fran</td>
<td>2-25</td>
<td>Data and Discrimination 1</td>
<td>Fran</td>
</tr>
<tr>
<td>3-1</td>
<td>Data and Discrimination 2</td>
<td>Fran</td>
<td>3-4</td>
<td>Data and Elections 1</td>
<td>Fran</td>
</tr>
<tr>
<td>3-8</td>
<td>Data and Elections 2</td>
<td>Fran</td>
<td>3-11</td>
<td>NO CLASS / WRITING DAY</td>
<td></td>
</tr>
<tr>
<td>3-15</td>
<td>Data and Astronomy</td>
<td>Alyssa Goodman</td>
<td>3-18</td>
<td>Data Science</td>
<td>Fran</td>
</tr>
<tr>
<td>3-22</td>
<td>Digital Humanities</td>
<td>Brett Bobley</td>
<td>3-25</td>
<td>Data Stewardship and Preservation</td>
<td>Fran</td>
</tr>
<tr>
<td>3-29</td>
<td>Data and the IoT</td>
<td>Fran</td>
<td>4-1</td>
<td>Data and Smart Farms</td>
<td>Rich Wolski</td>
</tr>
<tr>
<td>4-5</td>
<td>Data and Self-Driving Cars</td>
<td>Fran</td>
<td>4-8</td>
<td>Data and Ethics 1</td>
<td>Fran</td>
</tr>
<tr>
<td>4-12</td>
<td>Data and Ethics 2</td>
<td>Fran</td>
<td>4-15</td>
<td>Cybersecurity</td>
<td>Fran</td>
</tr>
<tr>
<td>4-19</td>
<td>Data and Dating</td>
<td>Fran</td>
<td>4-22</td>
<td>Data and Social Media</td>
<td>Fran</td>
</tr>
<tr>
<td>4-26</td>
<td>Tech in the News</td>
<td>Fran</td>
<td>4-29</td>
<td>Wrap-up / Discussion</td>
<td>Fran</td>
</tr>
<tr>
<td>5-3</td>
<td>NO CLASS</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Writing Assignment – Personal Essay 1
Grading – Personal Essay
(2 of these, 10 points each)

Grade Distribution

- Talks, 30
- Personal Essays, 20
- Choice, 10
- Briefing, 15
- Op-Ed, 15
- Part., 10

Fran Berman, Data and Society, CSCI 4370/6370
Personal Essay: You and the Pandemic

• 450-525 words / 11 point font / 10 points
• Send .docx to bermaf@rpi.edu before/by Sunday, February 7 at midnight.

• TOPIC: The Pandemic and You
Choose one of the topics below and tell a story that you’re comfortable sharing. (Only Fran will read this).

Pick 1:
• What is the best thing that’s happened to you because of the pandemic

or

• What is the worst thing that’s happened to you because of the pandemic
Writing Personal Narratives / Storytelling

• **GOAL:** Tell your audience (Fran/general public) an interesting (true) story on the assigned topic

• **PURPOSE:** Personal essays explore a *specific experience* and tell the story from *your point of view*. They may illustrate how a personal conflict, event, or experience left a lasting impression or how it changed your views or perspective.

• **TONE:** Can be more conversational than formal writing but should *establish you as an articulate and credible individual*.

• **FORMAT:**
  • **Introduction** -- Grab the reader and summarize your points
  • **Body** – main text that tells the story / provides information / explains and supports your points
  • **Conclusion** – may include a lesson, message, moral, take-away
Tips and Grading Rubric for Personal Narratives

TIPS

• Create an **outline** of the piece (don’t turn this in) before you write with the main points.

• Do **more than one draft** before turning your piece in.

• **Spell and grammar check** your piece

• Relevant statistics or facts should be cited and included as endnotes.

• Resources for writing personal essays:
  • [https://www.indeed.com/career-advice/career-development/how-to-write-a-personal-essay](https://www.indeed.com/career-advice/career-development/how-to-write-a-personal-essay)

GRADING RUBRIC

(10 points total)

• **5 points – content**
  • Is the story compelling?
  • Does the content comply to the personal essay format?

• **5 points – writing**
  • Is there a clear tone and narrative?
  • Is it well-written (English, grammar, spelling, flow)?
Lecture 2 – The Data-driven World

Modeling the World on the Internet

• *Your world* represented by the Internet
  • Recommender systems
  • Filter bubbles and echo chambers

• *You represented on the Internet*
  • Profiles and collected data
  • Inferred data

• Impacts of engaging in a modeled reality
Everything is a representation on the Internet

• How does the Internet model your world?
  • What you see on the Internet depends on who it perceives you to be
  • What you see on the Internet is optimized, often to maximize your value as a product

• How does the Internet model you?
  • You on the Internet = your profile
  • Profiles developed from collected data, acquired data, inferred data
  • Information may be out of context, incorrect, irrelevant, not managed by humans
Recommender Systems

- **Recommender systems** provide users with personalized product and information offerings based on perceived user preferences
  - Custom suggestions based on user characteristics and past behavior
  - *Self-reinforcing*; can narrow exposure, exacerbate degenerative feedback loops
  - Recommender systems used by a variety of services: Netflix, Amazon, Spotify, YouTube, Facebook, on-line dating sites, etc.

- **Algorithmic issue**: How to tailor outcomes but broaden exposure?

- **Social issue**: How do user’s interests evolve in the presence of recommender systems?
  - User interests may degenerate based on internal dynamics, recommender systems can slow down or accelerate this process
Recommender system methods

- **Collaborative filtering methods**
  - New recommendations based solely on past interactions (no additional information needed)
  - Can be used to detect similar users and/or similar items
  - May or may not utilize a model (memory-based vs. model-based)
  - Drawback: “cold start problem”

- **Content-based filtering methods**
  - New recommendations based on past interactions and additional information
  - User information parametrizes a model of preferences
Recommender systems provide both convenience and a limiting “nudge” to the user

- **“Nudge”** (behavioral science) – use of positive reinforcement and indirect suggestions as ways to influence the behavior and decision making of groups and individuals (e.g. putting fruit at eye level)
  - Nudges alter people’s behavior in a predictable way without forbidding any options or changing their economic incentives

- **Parameters that can be varied in recommender systems**
  - Items shown to the user
  - Number of times an item is shown to the user
  - Size of pool of potential items
  - Accuracy / inaccuracy of predictions (amount of noise)

- **Ways to avoid degeneracy in recommender systems (results from DeepMind research simulations):**
  - Show items only finitely many times.
  - Grow the candidate pool of items shown to the user

- **Note that users typically want a good recommender system but not a quickly degenerating one ....**
Netflix and data – a competitive advantage

- Netflix has a 90 second window to help users before they leave the platform for another service
- Algorithms produce $1B in revenue from customer retention
- 80% Netflix views come from recommendations
- Recommendation infrastructure:
  - 1300 clusters based on user’s viewing preferences
  - 2K taste groups
  - 7K shows and movies in the catalogue
  - Recommender systems use AI and machine learning to develop customized suggestions
Netflix Recommender System

• Data Netflix collects/acquires to characterize each user
  • Viewer interactions with Netflix services like viewer ratings, viewing history, etc.
  • Time of day, days of the week, location, device, etc. where a viewer watches (public info)
  • IP address (from device)
  • How long you’ve watched a show
  • Interactions with customer service
  • Web history (cookies, web beacons, advertising identifiers -- from browser)
  • Supplemental information such as demographic data, interest-based data, Internet browsing behavior (from third parties)
  • Name, email, payment method, telephone number, content rating, reviews (user provided)

• Netflix uses your data
  • To bucket you into a Taste Community
  • To parametrize its recommender system
  • To optimize audio and video encoding and adaptive bitrate selection
  • To figure out what artwork to show you
  • To parametrize user models for company purchases and creative projects
  • To guide advertising spend, advertising creative, and channel mix to identify new subscribers

• Netflix shares data “for limited purposes” with service providers, third-party companies tied to promotional offers with Netflix and law enforcement (upon request)

• Netflix doesn’t sell member information, ads to other companies or have third party developers providing applications on the platform
Ways the Internet limits your world – Filter bubbles and echo chambers

• **Echo chamber**: effect of a user’s interest being positively or negatively reinforced by repeated exposure to a certain item or category of items

• **Filter bubble**: when a user encounters only information and opinions that conform to and reinforce their own beliefs

• Filter bubbles and echo chambers limit the world represented to you on the Internet. Do they limit your world view or conveniently target useful information?
What’s newsworthy?

• **Media chooses what you will see.** When there are less choices, we are seeing “agenda compression” and news homogeneity.

• U.S. Media Collection – percent of stories in 32 popular publications from Media Cloud mentioning “coronavirus or COVID or Wuhan”: 1/1/20-3/25/20
What bias does your news source have?


Stories mentioning “coronavirus or COVID or Wuhan”: 1/1/20-3/25/20

Fran Berman, Data and Society, CSCI 4370/6370
The diversity of stories in the news is falling in general; big stories contribute

<table>
<thead>
<tr>
<th>Focus</th>
<th>Before</th>
<th>Later</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>The President</strong></td>
<td>Obama presidency: 1 in 10 stories</td>
<td>Trump Presidency: 1 in 4 Stories</td>
</tr>
<tr>
<td><strong>Sports</strong></td>
<td>7% of the stories in 2014</td>
<td>5.2% of the stories in 2019</td>
</tr>
<tr>
<td><strong>Protest, Black Lives Matter, uprising</strong></td>
<td>Before Michael Brown’s killing: 7.4%</td>
<td>After Michael Brown’s killing: 11.3%</td>
</tr>
<tr>
<td></td>
<td>1 week after George Floyd’s killing: 21.5%</td>
<td>2 weeks after George Floyd’s killing: 32.7%</td>
</tr>
<tr>
<td></td>
<td>Just after Breonna Taylor’s killing: Less</td>
<td>After George Floyd’s death: More (currently roughly 2.5%)</td>
</tr>
</tbody>
</table>
How does the media decide what to cover?

- **Chartbeat** documents what readers will click on.
- Social media platforms play an important role.
  - Algorithms govern which stories Facebook and other social media platforms amplify.
- Provocative stories more likely to provoke comments than information stories (e.g. politics, personal health decisions).

*Chartbeat provides information on users and uptake*
Other factors also influence what you see

- Tough economics:
  - fewer reporters producing less content; key local stories go unreported
- Consolidation of news sources
- High visibility media events shape subsequent media coverage

Image: https://billmoyers.com/story/media-consolidation-should-anyone-care/
Media Cloud: Structuring the Data

- **Media Cloud provides open source platform** with tools for audience analysis
  - *Explorer*: analysis of how digital news media covers your topic of interest
  - *Topic Mapper*: User topic-generator for a deeper dive into issues.
  - *Source Manager*: DB of print, broadcast and digital news collections

- Stories indexed from tens of thousands of publications and adds metadata to each story to assist media researchers

- Platform organizes data by topic – 400+ topical bins

- **Most popular stories trending to fewer bins**
  - 2013: 37 most popular bins contained 50% of the stories
  - 2019: 24 most popular bins contained 50% of the stories
  - 2020: 14 most popular bins contained 50% of the stories (so far) [includes public health, disease]
Media compression and news diversity

• Is media agenda compression bad?
  • Harder for a novel stories to break through dominance that Trump and coronavirus have on media agenda in 2020
  • Stories that break through must be really big: Michael Brown’s killing in Ferguson, BLM, protests
  • Creating an echo chamber effect

• Promoting more diverse news
  • News outlets should regularly monitor the diversity of topics in their new coverage

• What should their objectives be? Focusing primarily on Chartbeat-type analysis prioritizes popularity and diminishes diversity

• Chicken and egg problem / shared responsibility?
  • News outlets have to find a way to maintain a viable business model with more diverse content
  • Audiences need to proactively read more widely
Lecture 2 Resources 1


Lecture 2 Resources 2


• Partisanship, Propaganda, and Disinformation: Online Media and the 2016 U.S. Presidential Election, https://dash.harvard.edu/bitstream/handle/1/33759251/2017-08_electionReport_0.pdf

• Chartbeat website: https://chartbeat.com/products/dashboards/

• Media Cloud website: https://mediacloud.org/

What happens when you click agree?

New York Times

Terms of Service tell you under what conditions you may use a product or service.

Apple End User License Agreement prohibits using its products to develop nuclear weapons.

Amazon Terms of Service allows Cloud Computing to be used in a zombie apocalypse.
They can give companies permission to share or sell your data and/or restrict your rights

**Match.com (Match, OK Cupid, Tinder, etc.) shares information**
- With other users
- With service providers and partners
- With other Match Group businesses
- For corporate transactions (including restructuring, mergers, acquisitions)
- When required by law
- To enforce legal rights
- With your consent or at your request

“We may also share this information with other Match Group companies and third parties (notably advertisers) to develop and deliver targeted advertising on our services and on websites or applications of third parties, and to analyze and report on advertising you see. We may combine this information with additional non-personal information or personal information in hashed, non-human readable form collected from other sources.”

**AT&T Wireless Arbitration Clause:**
- “This agreement requires the use of arbitration on an individual basis to resolve disputes, rather than jury trials or class actions, and also limits the remedies available to you in the event of a dispute.”
The balance of power between consumers and tech companies is skewed

• Consumers are outgunned ...
  • Terms and conditions can stretch to 20,000 words
  • Terms are opaque and explanations filled with legalese
  • Studies estimate that understanding
  • Terms and Conditions may require 14 years of education and would require 76+ workdays for tech companies.

• The burden is on consumers
  • Products and services are often “take it or leave it”

https://www.google.com/books/editition/Julius_Caesar/Soh9UVlqRMC?hl=en&gbpv=1&dq=Julius+Caesar+play&printsec=frontcover
Sparsity of legislation allows companies to insert provisions for the company’s benefit, potentially at the consumers’ expense

- **Instagram** and **Twitter** can read your private messages.
- **Indeed, LinkedIn** and **YouTube** can keep data you’ve deleted.
- **Craigslist** can take credit for your content.
- **BBC, The Guardian, Indiegogo, Healthline, Adobe, Reddit, Zillow, Bose, Discord**, etc., etc. use your personal data to employ targeted third-party advertising
- **DoorDash** and **Lyft** ask users to agree that they are not delivery or transportation businesses (sheltering them from liability and allowing them to consider their drivers to be contractors)
Where is the law?

• Online contracts date to when software was sold in a box
  
  • Terms of service were considered agreed to when customers opened the “shrink wrap”
  
  • 1996 ruling upholds this notion, giving tech companies latitude to make terms and conditions hard to fathom or hard to find

• Things are starting to change:
  
  • Congress has prohibited companies from barring negative reviews
  
  • Senator Brown (D-Ohio) has introduced the Data Accountability and Transparency Act to Congress, which restricts collection and sharing of personal data
  
  • Other bills in Congress would allow stronger consumer protections and data privacy


https://www.nbcnews.com/business/consumer/yes-you-can-post-negative-online-review-says-congress-n693001
What more should be done?

• New legislation is needed to create rules that require greater transparency around changes to companies’ terms of service and clearer means by which customers agree to them:
  • Explanations in plain English
  • Make clear when policies change and what the changes are
  • Gear Terms and Conditions to the audience (children, general public, etc.)
  • Disclose the rules at regular intervals
  • Prohibition of automatic contract updates without user agreement

• Explore U.S. policy akin to EU’s General Data Protection Regulation “no take it or leave it” policy [A data controller may not refuse service to users who decline consent to processing that is not strictly necessary in order to use the service. (Article 7(4))]
The article needs to go farther ... Beyond better Terms and Conditions, Consumer protections are needed

- Improving Terms and Conditions partially kicks the can down the road

- Protections (not just understanding) needed for many of the things tech consumers find so egregious:
  - Lack of control with respect to the collection, use, deletion, sharing and sale of personal data
  - No options to “take it or leave it” policies for services and products that have become critical digital infrastructure
  - Limits on remediation options
References

• “What happens when you click agree?” New York Times

• Amazon Terms of Service, https://aws.amazon.com/service-terms/

• Apple Licensed Application End User License Agreement,

• Match.com Terms and Conditions,
  https://www.match.com/registration/privacystatement.aspx#Section6

• AT&T Wireless Customer Agreement,

• Cases, https://edit.tosdr.org/cases

• GDPR, Wikipedia,
  https://en.wikipedia.org/wiki/General_Data_Protection_Regulation
Annotated Presentation
(Don’t present like this!)
References


• Amazon Terms of Service, https://aws.amazon.com/service-terms/


• Match.com Terms and Conditions, https://www.match.com/registration/privacystatement.aspx#Section6


• Cases, https://edit.tosdr.org/cases


• (Fran’s opinion)
What happens when you click agree?

New York Times

Terms of Service tell you under what conditions you may use a product or service

Apple End User License Agreement prohibits using its products to develop nuclear weapons.

Amazon Terms of Service allows Cloud Computing to be used in a zombie apocalypse.
They can give companies permission to share or sell your data and/or restrict your rights

**Match.com (Match, OK Cupid, Tinder, etc.) shares information**

- With other users
- With service providers and partners
- With other Match Group businesses
- For corporate transactions (including restructuring, mergers, acquisitions)
- When required by law
- To enforce legal rights
- With your consent or at your request

“We may also share this information with other Match Group companies and third parties (notably advertisers) to develop and deliver targeted advertising on our services and on websites or applications of third parties, and to analyze and report on advertising you see. We may combine this information with additional non-personal information or personal information in hashed, non-human readable form collected from other sources.”

**AT&T Wireless Arbitration Clause:**

- “This agreement requires the use of arbitration on an individual basis to resolve disputes, rather than jury trials or class actions, and also limits the remedies available to you in the event of a dispute.”
The balance of power between consumers and tech companies is skewed

• Consumers are outgunned …
  • Terms and conditions can stretch to 20,000 words
  • Terms are opaque and explanations filled with legalese
  • Studies estimate that understanding
  • Terms and Conditions may require 14 years of education and would require 76+ workdays for tech companies.

• The burden is on consumers
  • Products and services are often “take it or leave it”

Fran Berman, Data and Society, CSCI 4370/6370
Sparsity of legislation allows companies to insert provisions for the company’s benefit, potentially at the consumers’ expense

- **Instagram** and **Twitter** can read your private messages.
- **Indeed, LinkedIn** and **YouTube** can keep data you’ve deleted.
- **Craigslist** can take credit for your content.
- **BBC, The Guardian, Indiegogo, Healthline, Adobe, Reddit, Zillow, Bose, Discord**, etc., etc. use your personal data to employ targeted third-party advertising
- **DoorDash** and **Lyft** ask users to agree that they are not delivery or transportation businesses (sheltering them from liability and allowing them to consider their drivers to be contractors)
Where is the law?

• Online contracts date to when software was sold in a box
  • Terms of service were considered agreed to when customers opened the “shrink wrap”
  • 1996 ruling upholds this notion, giving tech companies latitude to make terms and conditions hard to fathom or hard to find

• Things are starting to change:
  • Congress has prohibited companies from barring negative reviews
  • Senator Brown (D-Ohio) has introduced the Data Accountability and Transparency Act to Congress, which restricts collection and sharing of personal data
  • Other bills in Congress would allow stronger consumer protections and data privacy


https://www.nbcnews.com/business/consumer/yes-you-can-post-negative-online-review-says-congress-n693001
What more should be done?

• New legislation is needed to create rules that require greater transparency around changes to companies’ terms of service and clearer means by which customers agree to them:
  • Explanations in plain English
  • Make clear when policies change and what the changes are
  • Gear Terms and Conditions to the audience (children, general public, etc.)
  • Disclose the rules at regular intervals
  • Prohibition of automatic contract updates without user agreement

• Explore U.S. policy akin to EU’s General Data Protection Regulation “no take it or leave it” policy [A data controller may not refuse service to users who decline consent to processing that is not strictly necessary in order to use the service. (Article 7(4))]
The article needs to go father ... Beyond better Terms and Conditions, Consumer protections are needed

- Improving Terms and Conditions partially kicks the can down the road

- Protections (not just understanding) needed for many of the things tech consumers find so egregious:
  - Lack of control with respect to the collection, use, deletion, sharing and sale of personal data
  - No options to “take it or leave it” policies for services and products that have become critical digital infrastructure
  - Limits on remediation options
References


• Amazon Terms of Service, https://aws.amazon.com/service-terms/


• Match.com Terms and Conditions, https://www.match.com/registration/privacystatement.aspx#Section6


• Cases, https://edit.tosdr.org/cases

Presentation structure and grading
Presentation components and grading metrics

Presentation components (10 minutes presentation + 5 minutes Q&A):

• **Summary and main points**
  • What is the article about?
  • What are the main points/questions/issues described of the article?

• **What are the data issues?**
  • How is data used to support the article’s point of view?

• **Discussion**
  • What questions/issues arise from reading this article?

*Note: You may need to read additional publications, websites for your presentations*

Presentation Grading Metrics:

**Talk (5 pts):**
• Is the presentation compelling?
• Does the presentation tell an interesting story?
• Did the speaker use the timeframe effectively?

**Visuals (4 pts):**
• Are the slides well-organized and informative?
• Do the slides help tell the story?
• Are the slides visually interesting?
• Is the font readable, are images used to help convey the points, etc.?

**Content (4 pts):**
• Does the speaker understand the topic?
• Has the speaker leveraged appropriate additional materials as needed to support their presentation?

**Q&A (2 pts):**
• Is the speaker well prepared for questions? Can they respond to them articulately?
More about presentations

Do 2 of these, 15 points each

- You are responsible for ensuring that you sign up for 2 during the semester
- Presentation articles will be given in class and can be found on the class website
- Send your slides as a .pdf attachment (please include your name) to bermaf@rpi.edu 15 minutes before the beginning of the class in which you will present.

TIPS:

- Practice your presentation. Use this as an opportunity to become a better public speaker
- DO NOT SPEND MOST OF THE TALK READING YOUR NOTES
- Talk so your audience can hear you. Engage with your audience
- Be prepared for questions and have an intelligent form of “I don’t know”
- Use visuals and text to make your slides interesting. Don’t make the font too small.
Your Turn!

• Presentations for February 1

• Presentations for February 4

• Presentations for February 8

• Presentations for February 11
  • “We’re banning facial recognition. We’re missing the point.” New York Times, https://www.nytimes.com/2020/01/20/opinion/facial-recognition-ban-privacy.html (Josh M.)
  • “This site published every face from Parler’s Capitol riot videos”, Wired, https://www.wired.com/story/faces-of-the-riot-capitol-insurrection-facial-recognition/ (Nate S.)